Machine Learning - MT 2017
2. Mathematical Basics

Christoph Haase

University of Oxford

October 11, 2017

About this lecture

- No Machine Learning without rigorous mathematics
- This should be the most boring lecture
- Serves as reference for notation used throughout the course
- If there are any holes make sure to fill them sooner than later
- Attempt Problem Sheet 0 to see where you are standing

Outline

Today's lecture

- Linear algebra
- Calculus
- Probability theory

Linear algebra

We will mostly work in the real vector space:

- Scalar: single number $r \in \mathbb{R}$
- Vector: array of numbers $\mathbf{x} = (x_1, \ldots, x_D) \in \mathbb{R}^D$ of dimension D
- Matrix: two-dimensional array $\mathbf{A} \in \mathbb{R}^{m \times n}$ written as

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{bmatrix}$$

Linear algebra

We will mostly work in the real vector space:

- Scalar: single number $r \in \mathbb{R}$
- Vector: array of numbers $\mathbf{x} = (x_1, \ldots, x_D) \in \mathbb{R}^D$ of dimension D
- Matrix: two-dimensional array $\mathbf{A} \in \mathbb{R}^{m \times n}$ written as

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{bmatrix}$$

  - vector $\mathbf{x}$ is a $\mathbb{R}^{D \times 1}$ matrix
  - $\mathbf{A}_{i,j}$ denotes $a_{i,j}$
  - $\mathbf{A}_{i,:}$ denotes i-th row
  - $\mathbf{A}_{:,i}$ denotes i-th column
  - $\mathbf{A}^\mathsf{T}$ is the transpose of $\mathbf{A}$ such that $(\mathbf{A}^\mathsf{T})_{i,j} = \mathbf{A}_{j,i}$
  - symmetric if $\mathbf{A} = \mathbf{A}^\mathsf{T}$
  - $\mathbf{A} \in \mathbb{R}^{n \times n}$ is diagonal if $\mathbf{A}_{i,j} = 0$ for all $i \neq j$
  - $\mathbf{I}_n$ is the $n \times n$ diagonal matrix s.t. $(\mathbf{I}_n)_{i,i} = 1$

Operations on matrices

- Addition: $\mathbf{C} = \mathbf{A} + \mathbf{B}$ s.t. $\mathbf{C}_{i,j} = \mathbf{A}_{i,j} + \mathbf{B}_{i,j}$ with $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{R}^{m \times n}$
  - associative: $\mathbf{A} + (\mathbf{B} + \mathbf{C}) = (\mathbf{A} + \mathbf{B}) + \mathbf{C}$
  - commutative: $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$

Operations on matrices

- Addition: $\mathbf{C} = \mathbf{A} + \mathbf{B}$ s.t. $\mathbf{C}_{i,j} = \mathbf{A}_{i,j} + \mathbf{B}_{i,j}$ with $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{R}^{m \times n}$
  - associative: $\mathbf{A} + (\mathbf{B} + \mathbf{C}) = (\mathbf{A} + \mathbf{B}) + \mathbf{C}$
  - commutative: $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$
- Scalar multiplication: $\mathbf{B} = r \cdot \mathbf{A}$ s.t. $\mathbf{B}_{i,j} = r \cdot \mathbf{A}_{i,j}$

Operations on matrices

- Addition: $\mathbf{C} = \mathbf{A} + \mathbf{B}$ s.t. $\mathbf{C}_{i,j} = \mathbf{A}_{i,j} + \mathbf{B}_{i,j}$ with $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{R}^{m \times n}$
  - associative: $\mathbf{A} + (\mathbf{B} + \mathbf{C}) = (\mathbf{A} + \mathbf{B}) + \mathbf{C}$
  - commutative: $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$
- Scalar multiplication: $\mathbf{B} = r \cdot \mathbf{A}$ s.t. $\mathbf{B}_{i,j} = r \cdot \mathbf{A}_{i,j}$
- Multiplication: $\mathbf{C} = \mathbf{A} \cdot \mathbf{B}$ s.t.

$$\mathbf{C}_{i,j} = \sum_{1 \leq k \leq n} \mathbf{A}_{i,k} \cdot \mathbf{B}_{k,j}$$

  with $\mathbf{A} \in \mathbb{R}^{m \times n}, \mathbf{B} \in \mathbb{R}^{n \times p}, \mathbf{C} \in \mathbb{R}^{m \times p}$
  - associative: $\mathbf{A} \cdot (\mathbf{B} \cdot \mathbf{C}) = (\mathbf{A} \cdot \mathbf{B}) \cdot \mathbf{C}$
  - not commutative in general: $\mathbf{A} \cdot \mathbf{B} \neq \mathbf{B} \cdot \mathbf{A}$
  - distributive wrt. addition: $\mathbf{A} \cdot (\mathbf{B} + \mathbf{C}) = \mathbf{A} \cdot \mathbf{B} + \mathbf{A} \cdot \mathbf{C}$
  - $(\mathbf{A} \cdot \mathbf{B})^\mathsf{T} = \mathbf{B}^\mathsf{T} \cdot \mathbf{A}^\mathsf{T}$
  - $\mathbf{v}$ and $\mathbf{w}$ are orthogonal if $\mathbf{v}^\mathsf{T} \cdot \mathbf{w} = 0$

Eigenvectors, eigenvalues, determinant, linear independence, inverses

- $\mathbf{v} \in \mathbb{R}^n$ is an eigenvector of $\mathbf{A} \in \mathbb{R}^{n \times n}$ with eigenvalue $\lambda \in \mathbb{R}$ if $\mathbf{A} \cdot \mathbf{v} = \lambda \cdot \mathbf{v}$
- $\mathbf{A}$ is positive (negative) definite if all eigenvalues are strictly greater (smaller) than zero

Eigenvectors, eigenvalues, determinant, linear independence, inverses

- $\mathbf{v} \in \mathbb{R}^n$ is an eigenvector of $\mathbf{A} \in \mathbb{R}^{n \times n}$ with eigenvalue $\lambda \in \mathbb{R}$ if $\mathbf{A} \cdot \mathbf{v} = \lambda \cdot \mathbf{v}$
- $\mathbf{A}$ is positive (negative) definite if all eigenvalues are strictly greater (smaller) than zero
- Determinant of $\mathbf{A} \in \mathbb{R}^{n \times n}$ with eigenvectors $\lambda_1, \ldots, \lambda_n$ is

$$\det(\mathbf{A}) = \lambda_1 \cdot \lambda_2 \cdots \lambda_n$$

Eigenvectors, eigenvalues, determinant, linear independence, inverses

- $\mathbf{v} \in \mathbb{R}^n$ is an eigenvector of $\mathbf{A} \in \mathbb{R}^{n \times n}$ with eigenvalue $\lambda \in \mathbb{R}$ if $\mathbf{A} \cdot \mathbf{v} = \lambda \cdot \mathbf{v}$
- $\mathbf{A}$ is positive (negative) definite if all eigenvalues are strictly greater (smaller) than zero
- Determinant of $\mathbf{A} \in \mathbb{R}^{n \times n}$ with eigenvectors $\lambda_1, \ldots, \lambda_n$ is

$$\det(\mathbf{A}) = \lambda_1 \cdot \lambda_2 \cdots \lambda_n$$

- $\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)} \in \mathbb{R}^D$ are linearly independent if there are no $r_1, \ldots, r_n \in \mathbb{R} \setminus \{0\}$ such that

$$\sum_{1 \leq i \leq n} r_i \cdot \mathbf{v}^{(i)} = \mathbf{0}$$

Eigenvectors, eigenvalues, determinant, linear independence, inverses

- $\mathbf{v} \in \mathbb{R}^n$ is an eigenvector of $\mathbf{A} \in \mathbb{R}^{n \times n}$ with eigenvalue $\lambda \in \mathbb{R}$ if $\mathbf{A} \cdot \mathbf{v} = \lambda \cdot \mathbf{v}$

- $\mathbf{A}$ is positive (negative) definite if all eigenvalues are strictly greater (smaller) than zero

- Determinant of $\mathbf{A} \in \mathbb{R}^{n \times n}$ with eigenvectors $\lambda_1, \ldots, \lambda_n$ is

$$\det(\mathbf{A}) = \lambda_1 \cdot \lambda_2 \cdots \lambda_n$$

- $\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)} \in \mathbb{R}^D$ are linearly independent if there are no $r_1, \ldots, r_n \in \mathbb{R} \setminus \{0\}$ such that

$$\sum_{1 \leq i \leq n} r_i \cdot \mathbf{v}^{(i)} = \mathbf{0}$$

- $\mathbf{A} \in \mathbb{R}^{n \times n}$ invertible if there is $\mathbf{A}^{-1} \in \mathbb{R}^{n \times n}$ s.t.

$$\mathbf{A} \cdot \mathbf{A}^{-1} = \mathbf{A}^{-1} \cdot \mathbf{A} = \mathbf{I}_n$$

Eigenvectors, eigenvalues, determinant, linear independence, inverses

- $\mathbf{v} \in \mathbb{R}^n$ is an eigenvector of $\mathbf{A} \in \mathbb{R}^{n \times n}$ with eigenvalue $\lambda \in \mathbb{R}$ if $\mathbf{A} \cdot \mathbf{v} = \lambda \cdot \mathbf{v}$
- $\mathbf{A}$ is positive (negative) definite if all eigenvalues are strictly greater (smaller) than zero
- Determinant of $\mathbf{A} \in \mathbb{R}^{n \times n}$ with eigenvectors $\lambda_1, \ldots, \lambda_n$ is

$$\det(\mathbf{A}) = \lambda_1 \cdot \lambda_2 \cdots \lambda_n$$

- $\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)} \in \mathbb{R}^D$ are linearly independent if there are no $r_1, \ldots, r_n \in \mathbb{R} \setminus \{0\}$ such that

$$\sum_{1 \leq i \leq n} r_i \cdot \mathbf{v}^{(i)} = \mathbf{0}$$

- $\mathbf{A} \in \mathbb{R}^{n \times n}$ invertible if there is $\mathbf{A}^{-1} \in \mathbb{R}^{n \times n}$ s.t.

$$\mathbf{A} \cdot \mathbf{A}^{-1} = \mathbf{A}^{-1} \cdot \mathbf{A} = \mathbf{I}_n$$

- Note that:
  - $\mathbf{A}$ is invertible if rows of $\mathbf{A}$ are linearly independent
  - equivalently if $\det(\mathbf{A}) \neq 0$
  - If $\mathbf{A}$ invertible then $\mathbf{A} \cdot \mathbf{x} = \mathbf{b}$ has solution $\mathbf{x} = \mathbf{A}^{-1} \cdot \mathbf{b}$

Vector norms

Vector norms allow us to talk about the length of vectors

- The L^p norm of $\mathbf{v} = (v_1, \ldots, v_D) \in \mathbb{R}^D$ is given by

$$\|\mathbf{v}\|_p = \left(\sum_{1 \leq i \leq D} |v_i|^p \right)^{1/p}$$

Vector norms

Vector norms allow us to talk about the length of vectors

- The L^p norm of $\mathbf{v} = (v_1, \ldots, v_D) \in \mathbb{R}^D$ is given by

$$\|\mathbf{v}\|_p = \left(\sum_{1 \leq i \leq D} |v_i|^p \right)^{1/p}$$

- Properties of L^p (which actually hold for any norm):
  - $\|\mathbf{v}\|_p = 0$ implies $\mathbf{v} = \mathbf{0}$
  - $\|\mathbf{v} + \mathbf{w}\|_p \leq \|\mathbf{v}\|_p + \|\mathbf{w}\|_p$
  - $\|r \cdot \mathbf{v}\|_p = |r| \cdot \|\mathbf{v}\|_p$ for all $r \in \mathbb{R}$

Vector norms

Vector norms allow us to talk about the length of vectors

- The L^p norm of $\mathbf{v} = (v_1, \ldots, v_D) \in \mathbb{R}^D$ is given by

$$\|\mathbf{v}\|_p = \left(\sum_{1 \leq i \leq D} |v_i|^p \right)^{1/p}$$

- Properties of L^p (which actually hold for any norm):
  - $\|\mathbf{v}\|_p = 0$ implies $\mathbf{v} = \mathbf{0}$
  - $\|\mathbf{v} + \mathbf{w}\|_p \leq \|\mathbf{v}\|_p + \|\mathbf{w}\|_p$
  - $\|r \cdot \mathbf{v}\|_p = |r| \cdot \|\mathbf{v}\|_p$ for all $r \in \mathbb{R}$

- Popular norms:
  - Manhattan norm L^1
  - Eucledian norm L^2
  - Maximum norm L^∞ where $\|\mathbf{v}\|_\infty = \max_{1 \leq i \leq D} |v_i|$

Vector norms

Vector norms allow us to talk about the length of vectors

- The L^p norm of $\mathbf{v} = (v_1, \ldots, v_D) \in \mathbb{R}^D$ is given by

$$\|\mathbf{v}\|_p = \left(\sum_{1 \leq i \leq D} |v_i|^p \right)^{1/p}$$

- Properties of L^p (which actually hold for any norm):
  - $\|\mathbf{v}\|_p = 0$ implies $\mathbf{v} = \mathbf{0}$
  - $\|\mathbf{v} + \mathbf{w}\|_p \leq \|\mathbf{v}\|_p + \|\mathbf{w}\|_p$
  - $\|r \cdot \mathbf{v}\|_p = |r| \cdot \|\mathbf{v}\|_p$ for all $r \in \mathbb{R}$

- Popular norms:
  - Manhattan norm L^1
  - Eucledian norm L^2
  - Maximum norm L^∞ where $\|\mathbf{v}\|_\infty = \max_{1 \leq i \leq D} |v_i|$

- Vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^D$ are orthonormal if $\mathbf{v}$ and $\mathbf{w}$ are orthogonal and $\|\mathbf{v}\|_2 = \|\mathbf{w}\|_2 = 1$

Calculus

Functions of one variable $f : \mathbb{R} \to \mathbb{R}$

- First derivative:

$$f'(x) = \frac{d}{dx} f(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

  - $f'(x^*) = 0$ means that $f(x^*)$ is a critical or stationary point
  - Can be a local minimum, a local maximum, or a saddle point
  - Global minima are local minima x^* with smallest $f(x^*)$
  - Second derivative test to (partially) decide nature of critical point

Calculus

Functions of one variable $f : \mathbb{R} \to \mathbb{R}$

- First derivative:

$$f'(x) = \frac{d}{dx} f(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

  - $f'(x^*) = 0$ means that $f(x^*)$ is a critical or stationary point
  - Can be a local minimum, a local maximum, or a saddle point
  - Global minima are local minima x^* with smallest $f(x^*)$
  - Second derivative test to (partially) decide nature of critical point

- Differentiation rules:

$$\frac{d}{dx} x^n = n \cdot x^{n-1} \qquad \frac{d}{dx} a^x = a^x \cdot \ln(a) \qquad \frac{d}{dx} \log_a(x) = \frac{1}{x \cdot \ln(a)}$$

$$(f+g)' = f' + g' \qquad (f \cdot g)' = f' \cdot g + f \cdot g'$$

- Chain rule: if $f = h(g)$ then $f' = h'(g) \cdot g'$

Calculus

Functions of multiple variables $f : \mathbb{R}^m \to \mathbb{R}$

- Partial derivative of $f(x_1, \ldots, x_m)$ in direction x_i at $\mathbf{a} = (a_1, \ldots, a_m)$:

$$\frac{\partial}{\partial x_i} f(\mathbf{a}) = \lim_{h \to 0} \frac{f(a_1, \ldots, a_i + h, \ldots, a_m) - f(a_1, \ldots, a_i, \ldots, a_m)}{h}$$

Calculus

Functions of multiple variables $f : \mathbb{R}^m \to \mathbb{R}$

- Partial derivative of $f(x_1, \ldots, x_m)$ in direction x_i at $\mathbf{a} = (a_1, \ldots, a_m)$:

$$\frac{\partial}{\partial x_i} f(\mathbf{a}) = \lim_{h \to 0} \frac{f(a_1, \ldots, a_i + h, \ldots, a_m) - f(a_1, \ldots, a_i, \ldots, a_m)}{h}$$

- Gradient (assuming f is differentiable everywhere):

$$\nabla_{\mathbf{x}} f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_m} \right) \quad \text{s.t.} \quad \nabla_{\mathbf{x}} f(\mathbf{a}) = \left(\frac{\partial f}{\partial x_1}(\mathbf{a}), \ldots, \frac{\partial f}{\partial x_m}(\mathbf{a}) \right)$$

  - Points in direction of steepest ascent
  - Critical point if $\nabla_{\mathbf{x}} f(\mathbf{a}) = \mathbf{0}$

Calculus

Functions of multiple variables $f : \mathbb{R}^m \to \mathbb{R}$

- Partial derivative of $f(x_1, \ldots, x_m)$ in direction x_i at $\mathbf{a} = (a_1, \ldots, a_m)$:

$$\frac{\partial}{\partial x_i} f(\mathbf{a}) = \lim_{h \to 0} \frac{f(a_1, \ldots, a_i + h, \ldots, a_m) - f(a_1, \ldots, a_i, \ldots, a_m)}{h}$$

- Gradient (assuming f is differentiable everywhere):

$$\nabla_{\mathbf{x}} f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_m} \right) \quad \text{s.t.} \quad \nabla_{\mathbf{x}} f(\mathbf{a}) = \left(\frac{\partial f}{\partial x_1}(\mathbf{a}), \ldots, \frac{\partial f}{\partial x_m}(\mathbf{a}) \right)$$

  - Points in direction of steepest ascent
  - Critical point if $\nabla_{\mathbf{x}} f(\mathbf{a}) = \mathbf{0}$

Functions of multiple variables to vectors $\mathbf{f} : \mathbb{R}^m \to \mathbb{R}^n$:

- $\mathbf{f}$ given as $\mathbf{f} = (f_1, \ldots, f_n)$ with $f_i : \mathbb{R}^m \to \mathbb{R}$

- Jacobian $\mathbf{J}$ of $\mathbf{f}$ is an $n \times m$ matrix such that

$$\mathbf{J}_{i,j} = \frac{\partial f_i}{\partial x_j}$$

Calculus

Second-order derivatives of $f : \mathbb{R}^m \to \mathbb{R}$:

- Hessian is square matrix consisting of all second-order derivatives:

$$\mathbf{H}(f)(\mathbf{x})_{i,j} = \frac{\partial^2}{\partial x_i \partial x_j} f(\mathbf{x})$$

  - Symmetric (at continuous points)
  - If $\mathbf{H}(f)(\mathbf{a})$ positive (negative) definite then critical point $\mathbf{a}$ is local minimum (maximum)
  - Second derivative test may be inconclusive

Calculus

Second-order derivatives of $f : \mathbb{R}^m \to \mathbb{R}$:

- Hessian is square matrix consisting of all second-order derivatives:

$$\mathbf{H}(f)(\mathbf{x})_{i,j} = \frac{\partial^2}{\partial x_i \partial x_j} f(\mathbf{x})$$

  - Symmetric (at continuous points)
  - If $\mathbf{H}(f)(\mathbf{a})$ positive (negative) definite then critical point $\mathbf{a}$ is local minimum (maximum)
  - Second derivative test may be inconclusive

Useful differentiation rules:

$$\nabla_{\mathbf{x}}(\mathbf{c}^\mathsf{T}\mathbf{x}) = \mathbf{c}$$

$$\nabla_{\mathbf{x}}(\mathbf{x}^\mathsf{T}\mathbf{A} \cdot \mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{A}^\mathsf{T}\mathbf{x} \qquad (= 2\mathbf{A}\mathbf{x} \text{ for symmetric } \mathbf{A})$$

$$\nabla_{\mathbf{x}}(f + g) = \nabla_{\mathbf{x}} f + \nabla_{\mathbf{x}} g$$

$$\nabla_{\mathbf{x}}(f \cdot g) = f \cdot \nabla_{\mathbf{x}} g + g \cdot \nabla_{\mathbf{x}} f$$

Calculus

Second-order derivatives of $f : \mathbb{R}^m \to \mathbb{R}$:

- Hessian is square matrix consisting of all second-order derivatives:

$$\mathbf{H}(f)(\mathbf{x})_{i,j} = \frac{\partial^2}{\partial x_i \partial x_j} f(\mathbf{x})$$

  - Symmetric (at continuous points)
  - If $\mathbf{H}(f)(\mathbf{a})$ positive (negative) definite then critical point $\mathbf{a}$ is local minimum (maximum)
  - Second derivative test may be inconclusive

Useful differentiation rules:

$$\begin{aligned}
\nabla_{\mathbf{x}}(\mathbf{c}^{\mathsf{T}}\mathbf{x}) &= \mathbf{c} \\
\nabla_{\mathbf{x}}(\mathbf{x}^{\mathsf{T}}\mathbf{A}\cdot\mathbf{x}) &= \mathbf{A}\mathbf{x} + \mathbf{A}^{\mathsf{T}}\mathbf{x} \qquad (= 2\mathbf{A}\mathbf{x} \text{ for symmetric } \mathbf{A}) \\
\nabla_{\mathbf{x}}(f + g) &= \nabla_{\mathbf{x}} f + \nabla_{\mathbf{x}} g \\
\nabla_{\mathbf{x}}(f \cdot g) &= f \cdot \nabla_{\mathbf{x}} g + g \cdot \nabla_{\mathbf{x}} f
\end{aligned}$$

See `http://en.wikipedia.org/wiki/Matrix_calculus` for many more useful rules, and use them!

Chain rule in higher dimensions

Let $\mathbf{y} = g(\mathbf{x})$, $z = f(\mathbf{y})$ for $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^n$:

$$\frac{\partial z}{\partial x_i} = \sum_j \frac{\partial z}{\partial y_j} \cdot \frac{\partial y_j}{\partial x_i}$$

$$\nabla_{\mathbf{x}} z = \mathbf{J}_g^{\mathsf{T}} \cdot \nabla_{\mathbf{y}} z = \frac{\partial \mathbf{y}}{\partial \mathbf{x}} \cdot \nabla_{\mathbf{y}} z$$

Chain rule in higher dimensions

Let $\mathbf{y} = g(\mathbf{x})$, $z = f(\mathbf{y})$ for $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^n$:

$$\frac{\partial z}{\partial x_i} = \sum_j \frac{\partial z}{\partial y_j} \cdot \frac{\partial y_j}{\partial x_i}$$

$$\nabla_{\mathbf{x}} z = \mathbf{J}_g^{\mathsf{T}} \cdot \nabla_{\mathbf{y}} z = \frac{\partial \mathbf{y}}{\partial \mathbf{x}} \cdot \nabla_{\mathbf{y}} z$$

Example

Let $g(x, y) = (x^2, y)$, $f(s, t) = (s + t)^2$ and $z = f(g(x, y))$. Then

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial s} \cdot \frac{\partial s}{\partial x} + \frac{\partial z}{\partial t} \cdot \frac{\partial t}{\partial x} = 2 \cdot (x^2 + y) \cdot 1 \cdot 2 \cdot x + 2 \cdot (x^2 + y) \cdot 1 \cdot 0 = 4x(x^2 + y)$$

$$\mathbf{J}_g^{\mathsf{T}} = \begin{bmatrix} 2 \cdot x & 0 \\ 0 & 1 \end{bmatrix}$$

$$\nabla_{\mathbf{y}} z = (2 \cdot (x^2 + y), 2 \cdot (x^2 + y))$$

$$\nabla_{\mathbf{x}} z = (4 \cdot x \cdot (x^2 + y), 2 \cdot (x^2 + y))$$

Probability theory

Probability space:

- Consists of sample space S and a probability function $p : \mathcal{P}(S) \to [0, 1]$ assigning a probability to every event
- Fulfills axioms of probability:
  - $p(\emptyset) = 0$ and $p(S) = 1$
  - For mutually exclusive events $A_1, A_2, \ldots$

$$p\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} p(A_i)$$

Probability theory

Probability space:

- Consists of sample space S and a probability function $p : \mathcal{P}(S) \to [0, 1]$ assigning a probability to every event
- Fulfills axioms of probability:
  - $p(\emptyset) = 0$ and $p(S) = 1$
  - For mutually exclusive events $A_1, A_2, \ldots$

$$p\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} p(A_i)$$

Trivial properties:

- $p(\overline{A}) = 1 - p(A)$
- If $A \subseteq B$ then $p(A) \leq p(B)$
- $p(A \cup B) = p(A) + p(B) - p(A \cap B)$

Probability theory

Conditional probability:

- Given events A, B with $p(B) > 0$, conditional probability of A given B is

$$p(A|B) = \frac{p(A \cap B)}{p(B)}$$

- $p(A)$ is prior, and $p(A|B)$ is posterior probability of A
- Law of total probability: Given partition $A_1, \ldots, A_n$ of S with $p(A_i) > 0$,

$$p(B) = \sum_{i=1}^{n} p(B|A_i) \cdot p(A_i)$$

- Bayes' rule:

$$p(A|B) = \frac{p(B|A) \cdot p(A)}{p(B)}$$

Probability Theory

Random variable (r.v.):

- Function from sample space to some numeric domain (usually $\mathbb{R}$)
- $p(X = x)$ denotes probability of event $\{s \in S : X(s) = x\}$
- Write $X \sim p(x)$ to specify probability distribution of X

Probability Theory

Random variable (r.v.):

- Function from sample space to some numeric domain (usually $\mathbb{R}$)
- $p(X = x)$ denotes probability of event $\{s \in S : X(s) = x\}$
- Write $X \sim p(x)$ to specify probability distribution of X

Discrete random variables:

- Discrete if there are $a_1, a_2, \ldots$ such that $p(X = a_j \text{ for some } j) = 1$
- Probability mass function (PMF) p_X given by $p_X(x) = p(X = x)$ giving distribution of X
- Cumulative distribution function (CDF) maps x to $p(X \leq x)$

Probability Theory

Random variable (r.v.):

- Function from sample space to some numeric domain (usually $\mathbb{R}$)
- $p(X = x)$ denotes probability of event $\{s \in S : X(s) = x\}$
- Write $X \sim p(x)$ to specify probability distribution of X

Discrete random variables:

- Discrete if there are $a_1, a_2, \ldots$ such that $p(X = a_j \text{ for some } j) = 1$
- Probability mass function (PMF) p_X given by $p_X(x) = p(X = x)$ giving distribution of X
- Cumulative distribution function (CDF) maps x to $p(X \leq x)$

Continuous random variables:

- Continuous if CDF is differentiable
- Probability density function (PDF) $p(x)$ is derivative of CDF giving distribution of X

Probability Theory

Joint probability distributions:

- Natural generalisation to vectors of random variables giving joint probability distributions, e.g., $p(X = x, Y = y)$
- Marginal probability distribution: Given $p(X, Y)$, obtain $p(X)$ via

$$p(X = x) = \sum_y p(X = x, Y = y) \qquad \text{resp.} \qquad p(x) = \int p(x, y)\, dy$$

- Conditional probabilities: Assuming $p(X = x) > 0$,

$$p(Y = y \mid X = x) = \frac{p(Y = y, X = x)}{p(X = x)}$$

- Chain rule of conditional probability:

$$p(X^{(1)}, \ldots, X^{(n)}) = p(X^{(1)}) \cdot \prod_{i=2}^{n} p(X^{(i)} \mid X^{(1)}, \ldots, X^{(i-1)})$$

Probability Theory

Expected value of random variable w.r.t. f:

- $\mathbb{E}_{X\sim p}[f(x)] = \sum_x p(x) \cdot f(x)$ (for discrete r.v.'s)
- $\mathbb{E}_{X\sim p}[f(x)] = \int p(x) \cdot f(x)\, dx$ (for continuous r.v.'s)
- Linearity of expectation:

$$\mathbb{E}_X[\alpha \cdot f(x) + \beta \cdot g(x)] = \alpha \cdot \mathbb{E}_X[f(x)] + \beta \cdot \mathbb{E}_X[g(x)]$$

Probability Theory

Expected value of random variable w.r.t. f:

- $\mathbb{E}_{X\sim p}[f(x)] = \sum_x p(x)\cdot f(x)$ (for discrete r.v.'s)
- $\mathbb{E}_{X\sim p}[f(x)] = \int p(x)\cdot f(x)\,dx$ (for continuous r.v.'s)
- Linearity of expectation:

$$\mathbb{E}_X[\alpha\cdot f(x) + \beta\cdot g(x)] = \alpha\cdot\mathbb{E}_X[f(x)] + \beta\cdot\mathbb{E}_X[g(x)]$$

Properties of random variables:

- Variance captures how much values of probability distribution vary on average if randomly drawn:

$$\mathrm{Var}(f(x)) = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2]$$

- Standard deviation is square root of variance

$$\mathrm{SD}(f(X)) = \sqrt{\mathrm{Var}(f(x))}$$

- Covariance generalises variance to two r.v.'s:

$$\mathrm{Cov}(f(x), g(y)) = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])\cdot(g(y) - \mathbb{E}[g(y)])]$$

- Covariance matrix $\mathbf{\Sigma}$ generalises covariance to multiple r.v.'s x_i:

$$\mathbf{\Sigma}_{i,j} = \mathrm{Cov}(f_i(x_i), f_j(x_j))$$

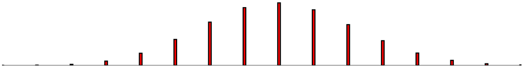
Well-known discrete probability distributions:

- Bernoulli:
  - Parameter: $\phi \in [0, 1]$
  - PMF: $p(X = 1) = \phi$, $p(X = 0) = 1 - \phi$;
  - $\mathbb{E}[X] = \phi$; $\mathrm{Var}(X) = \phi \cdot (1 - \phi)$

Probability Theory

Well-known discrete probability distributions:

- Bernoulli:
  - Parameter: $\phi \in [0,1]$
  - PMF: $p(X=1) = \phi$, $p(X=0) = 1-\phi$;
  - $\mathbb{E}[X] = \phi$; $\mathrm{Var}(X) = \phi \cdot (1-\phi)$
- Binomial distribution:
  - Parameters: $\phi \in [0,1]$, $n \in \mathbb{N} \setminus \{0\}$
  - PMF: $p(X=k) = \binom{n}{k} \cdot \phi^k \cdot (1-\phi)^{n-k}$
  - $\mathbb{E}[X] = n \cdot \phi$; $\mathrm{Var}(X) = n \cdot \phi \cdot (1-\phi)$

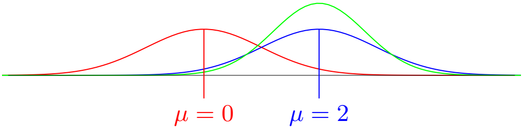


Well-known continuous probability distributions:

- Normal distribution:
  - Parameters: μ, σ^2
  - PDF:

$$\mathcal{N}(x;\mu,\sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

  - $\mathbb{E}[X] = \mu$; $\mathrm{Var}(X) = \sigma^2$

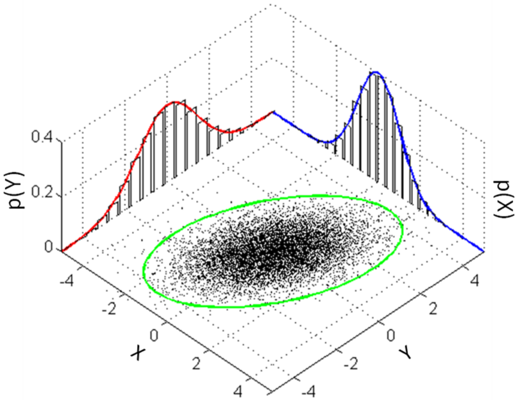


Probability Theory

- Multivariate normal distribution:
  - Parameters: $k, \mu, \mathbf{\Sigma}$ positive semi-definite
  - PDF:

$$\mathcal{N}(\mathbf{x}; \mu, \mathbf{\Sigma}) = \sqrt{\frac{1}{(2\pi)^k \det(\mathbf{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\mathsf{T}\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right)$$

  - $\mathbb{E}[\mathbf{X}] = \mu$; $\mathrm{Var}(\mathbf{X}) = \mathbf{\Sigma}$

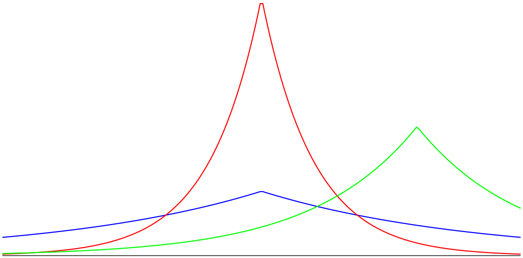


Well-known continuous probability distributions:

- Laplace distribution:
  - Parameters: μ, γ^2
  - PDF:

$$\mathrm{Lap}(x; \mu, \gamma) = \frac{1}{2\gamma} \exp\left(-\frac{|x-\mu|}{\gamma}\right)$$

  - $\mathbb{E}[X] = \mu$; $\mathrm{Var}(X) = 2\gamma^2$



Next Time

- Supervised Machine Learning: Linear regression